UCRL-MI-216970

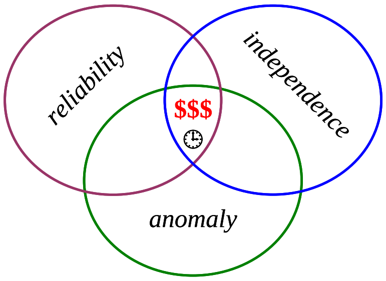
Utility of Information in Data through Relative-Value Ranking[*]

Tina Eliassi-Rad

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
7000 East Ave, L-560, Livermore, CA 94550
eliassirad1@llnl.gov (925)422-1552

Abstract

The intelligence community receives tens of tera-bytes of data per day [8]. Even with our fastest computer systems, it is difficult not to be overwhelmed by the sheer amount of data. However, data is only interesting with respect to the information that it carries. Namely, some pieces of information are more valuable than others. We propose to develop a mathematical model that calculates a relative-value ranking of information in data. In particular, we will build a formal model that ascertains the value of a piece of information based on three factors: *reliability of source*, *independence between sources*, and *contextual fit with respect to prior knowledge*. Our formal model will allow analysts to gain actionable knowledge in a variety of ways, such as (*i*) recognizing propaganda in data sources (e.g., by detecting deception through sudden or dramatic changes in the value of information from a data source), (*ii*) testing a hypothesis (e.g., by measuring the quality and quantity of information needed to either confirm or disconfirm the hypothesis), (*iii*) influencing the collection of data (e.g., by changing the way data is collected at sources with lower informativeness), and (*iv*) allowing triage on the massive amounts of data (e.g., by processing data sources with higher informativeness first). Moreover, our approach will incorporate the temporal dimension to address the diminishing marginal returns on information that arrives late. In short, our model will enable us to capture and rank *utility of information in data with respect to time and a subject matter*.



1. Background and Motivation

To our best knowledge, no one has tried to develop a formal model that ascertains the value of a piece of information and ranks it among other pieces of information (within the temporal dimension and with respect to any given subject matter such as Avian influenza or a taxonomy such as viral infections). The most relevant work has been in philosophy of science [1][2][4][5]. Bovens and Hartmann [1] consider reliability and contextual fit within the framework of a Bayesian network, which sidesteps the problem of independence between sources. In [2], Bovens and Hartmann address the issue of variety of evidence in support of a given hypothesis, which is different from determining whether two sources are independent of each other. Fitelson [4] provides a Bayesian account for measuring independence across various sources. In other work, Fitelson [5] studies the degree of confirmation for a specific hypothesis given a new piece of information and prior knowledge. However, Fitelson does not consider reliability or contextual fit in either [4] or [5]. Work on reliability has mostly been in the context of information fusion. Noble [7] provides a nice overview of the reliability of open source information. However, he does not formally address confirmational independence and contextual fit. Our work will be the first of its kind in terms of measuring utility of information from real-world (dynamic) data sources. Such quantitative metrics will be extremely useful in producing accurate actionable knowledge [6][9].

2. Proposed Work and Technical Approach

The goal of this proposed work is to understand how to mathematically measure utility of information. To achieve this goal, we first need to define what constitutes as information. We envision a piece of information to be the smallest amount of evidence needed to update either a database of facts or the probability distribution of a hypothesis (within a database of hypotheses). Naturally, information is conditional on a chosen domain topic (e.g., Avian influenza) or a selected taxonomy (e.g., viral infections).

We also need to define a consistent framework for objectively assigning values to reliability of source, confirmational independence, and contextual fit with respect to a given subject matter and time. Temporal tracking of information is very important since the value of information is strongly dependent on when the information is received (see Section 2.2 for more details on this).

To assess the reliability of a source, we will build statistical models that use frequencies of true positive and false positive information originating for the given data source. We imagine that there exists plenty of information that does not fall in either the true positive or the false positive categories. In such cases, we will use clustering (i.e. grouping) algorithms to develop a spectrum around the two categories. We will compute confirmational independence by measuring the degree to which pieces of information from two different sources confirm each other. This degree of confirmation will allow us to statistically capture how independent two sources are with respect to their contents. Finally, we will calculate the degree of contextual fit of a new piece of information to our (*a priori*) belief set by using probabilistic functions for belief expansion and revision. These functions will capture the amount of disparity in our beliefs given the new piece of evidence. Contextual fit and reliability exhibit an inverse relationship [1]. That is, a piece of information from a less reliable source must have a higher contextual fit to be considered for belief expansion (as opposed to a piece of information from a more reliable source which can afford to have a lower contextual fit). Given objective statistical approaches for measuring reliability, independence, and contextual fit, we face the following technical challenges.

2.1 Challenge: Data at Various Manipulation Levels

We understand that intelligence data comes at various levels of manipulation. In particular, we envision the following five levels:

Levels	Category	Description
1	Raw	Data is collected at the source without any manipulation.
2	Calibrated	Data is reduced through the process of feature selection.
3	Interpreted	Data from the calibrated level is "interpreted" by a domain expert (such as an infectious diseases analyst). The feature selection at this level is more semantically oriented.
4	Extracted	Data from the interpreted level is put in context with previously extracted data.
5	Exploited	Data is converted into an "action report," where decisions are made.

It is seldom possible to obtain raw data. We will construct our model to work with two kinds of data: (*i*) *filtered* data that falls under the categories of calibrated or interpreted data; and (*ii*) *finished* data which covers the categories of extracted or exploited data.

As data get manipulated from one level to the next, biases get introduced. These biases must be accounted for. Our utility model can easily measure such biases by highlighting the difference between utilities of information from different levels.

2.2 Challenge: Temporal Tracking

Utility of information strongly depends on when it is received. Information like any other commodity follows the law of diminishing marginal returns. Therefore, any approach that measures utility of information must penalize information that is saturated and/or arrives too late (namely, the "overcome by events" phenomenon). A "timely" piece of information dramatically changes your prior beliefs and so "does not fit within your context" (assuming we hold the independence and reliability factors fixed). The less a piece of information contributes to increasing your belief set (with respect to facts and hypotheses), the less timely it is. For example, information about a potential World Trade Center attack that arrived on September 12, 2001 was too late! In this scenario, you are getting information about an event already in your facts database. Tracking the temporal dimension within our model decreases the time period required to form actionable knowledge and thus improves the timeliness of intelligence.

2.3 Challenge: Learning a Utility Function

The parameters into our utility function (that computes a relative-value ranking for different pieces of information) are: reliability of source, confirmational independence, contextual fit, a given domain topic, and the temporal dimension. Combining these parameters into an appropriate function is not an easy task. For example, how much weight should reliability of source have in the utility function; how does this weight compare to the weights put on conformational independence and contextual fit? Moreover, the price value of information will depend on the amount to which it contributes to either the facts database or the hypotheses database. What is the function that captures the relationship between utility of information and contributions to these two databases; what are the properties of such a function? We envision a machine learning approach for this task. Specifically we will study algorithms that learn an expected utility function [3] for information relating to a given subject matter (over the temporal dimension) by using measurements from information theory (such as information gain).

3. Deliverables and Milestones

We will demonstrate the applicability of utility of information in real-world applications by evaluating our approach on data from the World-Wide Web. Data from the Web is ideal for our empirical studies because it mimics the real-world very well. Web data is abundant (about 10 tera-bytes per day), dynamic (as in continually changing), and noisy. It comes from sites that are not independent of each other, sites that want to spread propaganda and deceive, *etc*. For our historical data, we will use the Internet Archive at http://www.archive.org.

In addition to regular status reports, we will deliver technical reports describing our approach, summarize the results of our investigation, and offer lessons learned for future research. Methods and results obtained using unclassified sources will be submitted for peer review in appropriate professional venues. Redistribution of the source data harvested from the Internet Archive will be subject to the terms of use associated with that data. Any software developed specifically for this project will be freely provided in source form, though The University of California will retain all intellectual property of the software and research.

Our first year milestones are:

- Develop and validate algorithms for independently measuring reliability, independence, and contextual fit.
- Implement a testbed from the Web and the Internet Archive.
- Empirically investigate the interactions between reliability, independence, and contextual fit on our testbed; publish a report on the results of the experiments.
- Investigate and prototype various techniques for combining reliability, independence, and contextual fit to measure the utility of and rank different pieces of information.

In our second and third years, we will continue research, development, and evaluation of various techniques for measuring utility of information with an emphasis on hypothesis generation, representation, and testing. In particular, we will focus on algorithms that use our utility model to transfer information from the hypotheses database to the facts database and vice versa. Finally, we will investigate approaches for combining utility functions from complimentary subject matters that belong to the same taxonomy (e.g., Avian influenza and the Spanish pandemic flu of 1918 both fall under the taxonomy of viral infections). Simply averaging these utility functions won't work due to Simpson's paradox [10].

4. Qualifications of the Principal Investigator and Research Team

The principal investigator, **Tina Eliassi-Rad**, has been a computer scientist at the Center for Applied Scientific Computing (CASC) at Lawrence Livermore National Laboratory since 2001. Dr. Eliassi-Rad is an expert in machine learning, artificial intelligence, computational statistics, knowledge discovery and data mining. She has worked on analysis of various large-scale data sets such the World Wide Web, scientific simulation data, and heterogeneous complex networks. Dr. Eliassi-Rad will be the project's liaison with collaborators and will be involved in all algorithmic research and prototyping.

Mr. Brian Gallagher is a computer scientist at the Center for Applications Development and Software Engineering (CADSE) at Lawrence Livermore National Laboratory. He has extensive experience in designing, implementing and conducting experiments in relational knowledge discovery and communication networks. Mr. Gallagher will be involved in algorithmic research and development.

5. Collaborations

Our collaborators are **Professor Branden Fitelson** at University of California at Berkeley (http://fitelson.org) and **Professor Dan Roth** at University of Illinois at Urbana-Champaign (http://l2r.cs.uiuc.edu/~danr/). Professor Fitelson is an expert on confirmation theory, probability theory, decision theory, game theory, philosophy of science, logic, and epistemology. His 2001 doctoral dissertation titled "Studies in Bayesian Confirmation Theory" is a seminal work in the field of confirmation theory [5]. Our collaboration with Professor Fitelson will focus on the mathematical foundations of independent evidence and degree of confirmation.

Professor Dan Roth is an expert in machine learning, knowledge representation, and reasoning. His work on performing knowledge intensive inference has been influential in analysis of noisy data. Professor Roth is familiar with the IC community's challenges having received previous funding from ARDA. Our collaboration with Professor Roth will focus on both algorithmic and empirical studies of our proposed approach.

6. Budget

To meet our deliverables for phase 1, we will need approximately $450,000 for twelve months. This budget will support Eliassi-Rad (50%), Brian Gallagher (50%), and our collaboration with Professor Roth (namely, to support a graduate student's research on this proposal). Professor Fitelson has agreed to participate on this project with his own funds.

References

[1] Luc Bovens and Stephan Hartmann, Belief Expansion, Contextual Fit, and the Reliability of Information Sources, In *Proceedings of Modeling and Using Context*: *Third International and Interdisciplinary Conference* (Dundee, UK; July 2001), Springer.

[2] Luc Bovens and Stephan Hartmann, *Bayesian Epistemology*, Oxford University Press, 2003.

[3] Morton Davis, *Game Theory: A Nontechnical Introduction*, Dover, 1983.

[4] Branden Fitelson, A Bayesian Account of Independent Evidence with Applications, In *Proceedings of the Philosophy of Science*, vol. 68, 2001, pp. S123–S140.

[5] Branden Fitelson, *Studies in Bayesian Confirmation Theory*, Ph.D. Dissertation, University of Wisconsin-Madison, 2001.

[6] Sherman Kent, Words of Estimative Probability, In *Studies in Intelligence* (Fall 1964), http://www.cia.gov/csi/books/shermankent/6words.html.

[7] David Noble, Assessing the Reliability of Open Source Information, In *Proceedings of the Seventh International Conference on Information Fusion* (Stockholm, Sweden; June 2004), pp. 1172–1178.

[8] Michael Pagels, Avoiding Death by Data, In *Proceedings of DARPATech 2005* (Anaheim, CA; August 2005), pp. 98–101.

[9] Michael Schrage, What Percent Is 'Slam Dunk'? Give Us Odds on Those Estimates, In *Washington Post* (February 20, 2005), p. B01.

[10] E. H. Simpson, The Interpretation of Interaction in Contingency Tables, In *Journal of the Royal Statistical Society*, series B, vol. 13, 1951, pp. 238-241.

* This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48. UCRL-MI-216970. We thank Brett Wayne, Dan Roth, and Branden Fitelson for valuable discussions.